

Substrate Recognition Sites in Cytochrome P450 Family 2 (CYP2) Proteins Inferred from Comparative Analyses of Amino Acid and Coding Nucleotide Sequences*

(Received for publication, July 23, 1991)

Osamu Gotoh†

From the Department of Biochemistry, Saitama Cancer Center Research Institute, 818 Komuro, Ina-machi, Saitama 362, Japan

The substrate recognition regions in cytochrome P450 family 2 (CYP2) proteins were inferred by group-to-group alignment of CYP2 sequences and those of bacterial P450s, including *Pseudomonas putida* P450 101A (P450_{cam}), whose substrate-binding residues have been definitely identified by x-ray crystallography of a substrate-bound form (Poulos T. L., Finzel, B. C., and Howard, A. J. (1987) *J. Mol. Biol.* 195, 687-700). The six putative substrate recognition sites, SRSs, thus identified are dispersively located along the primary structure and constitute about 16% of the total residues. All the reported point mutations and chimeric fragments that significantly affect the substrate specificities of the parental CYP2 enzymes fell within or overlapped some of the six SRSs. Analysis of nucleotide substitution patterns in closely related members in four subfamilies, CYP2A, 2B, 2C, and 2D, consistently indicated that the SRSs have accumulated more nonsynonymous (amino acid-changing) substitutions than the rest of the sequence. This observation supports the idea that diversification of duplicate genes of drug-metabolizing P450s occurs primarily in substrate recognition regions to cope with an increasing number of foreign compounds.

The cytochrome P450 (P450)¹ monooxygenase system plays a central role in the metabolism of a wide variety of foreign compounds such as plant metabolites, environmental pollutants, and drugs. The human and rodent genomes contain at least 50 P450 genes, which are classified into 10 families according to the currently available protein sequence data (Nebert *et al.*, 1991). Enzymes belonging to families 1² through 4 are mainly concerned with catabolism of thousands of endogenous and exogenous chemicals, whereas members of other mammalian P450 families catalyze specific reactions involved in physiologically important pathways of synthesis of steroid hormones, prostaglandins, and vitamin D₃ (Gonzalez, 1990). The exact number of active P450 genes in a mammalian genome which encode drug-metabolizing P450s is not known but may not greatly exceed 50. These P450

enzymes show partial overlap but distinct substrate specificities. Thus it is of great interest to elucidate the molecular mechanisms underlying the broad but specific metabolic capacities of the mammalian P450 systems consisting of relatively limited numbers of catalysts.

A primary question is which parts of a P450 protein are involved in recognition or binding of substrates and hence determine the substrate specificity. There have been various experimental and analytical investigations on this problem, including chemical modifications with substrate analogues (Onoda *et al.*, 1987), site-directed mutagenesis (Imai and Nakamura, 1989; Aoyama *et al.*, 1989; Lindberg and Negishi, 1989; Matsunaga *et al.*, 1990b; Zhou *et al.*, 1991), protein engineering with chimeric constructs (Sakaki *et al.*, 1987; Imai, 1988; Pompon and Nicolas, 1989; Uno and Imai, 1989; Kronbach *et al.*, 1989; Uno *et al.*, 1990), searching for similar subsequences in other protein families with similar substrate or binding specificities (Gotoh *et al.*, 1985; Picado-Leonard and Miller, 1988), and sequence alignment (Gotoh and Fujii-Kuriyama, 1989; Laughton *et al.*, 1990) with P450 101A, the sole P450 whose substrate-binding residues have been definitely identified from the three-dimensional structure (Poulos *et al.*, 1985, 1987). The results of these studies have been controversial, and no unified view about the substrate recognition sites in mammalian P450s has been established.

One reason for this confusion arises from the difficulty in aligning distantly related protein sequences of a mammalian P450 and bacterial P450 101A. In fact, a mammalian P450 and P450 101A show only 12-20% identity of amino acids (Nelson and Strobel, 1987; Gotoh and Fujii-Kuriyama, 1989), and the typical alignment scores of 3.0-5.0 S.D. (Gotoh *et al.*, 1983; Gotoh and Fujii-Kuriyama, 1989) indicate that only marginal accuracy can be expected in pairwise alignments (Barton and Sternberg, 1987). However, a large number of mammalian P450 sequences are now available, and those of members of a given family are easily aligned. Several bacterial P450s with similar sequences to P450 101A have also been reported (Nebert *et al.*, 1991). Hence, it is possible to improve the accuracy of alignment between mammalian and bacterial sequences by a group-to-group comparison. Further improvement in the accuracy should be achieved by taking into consideration not only primary structures but also some properties associated with higher order structures such as secondary structure prediction (Garnier *et al.*, 1978; Gibrat *et al.*, 1987) and hydropathy indices (Kyte and Doolittle, 1982). Taking these properties into consideration, I have modified our basic alignment algorithms (Gotoh, 1982, 1990). This paper presents potential substrate recognition sites in mammalian P450s identified using an amended alignment.

The versatility of the protective cytochrome P450 systems against foreign compounds is reminiscent of those of globulin

* This work was partly supported by a grant-in-aid for scientific research on priority areas P450 from the Ministry of Education, Science, and Culture of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† Tel.: 81-48-722-1111; Fax: 81-48-722-1739.

¹ The abbreviations used are: P450, cytochrome P450; SRS, substrate recognition site.

² The recommendation of Nebert *et al.* (1991) for notations of P450 proteins and gene families was followed.

families, such as immunoglobulin, T-cell receptor, and major histocompatibility complex. On analysis of nucleotide substitution patterns in highly polymorphic major histocompatibility complex Class I and Class II genes, Hughes and Nei (1988, 1989) found that nonsynonymous (amino acid-replacing) codon changes occur at high rates within the antigen recognition site (Bjorkman *et al.*, 1987). A similar concentration of nonsynonymous codon changes has also been recognized in regions corresponding to antigen-binding sites in immunoglobulin genes (Tanaka and Nei, 1989). These observations are interpreted as a result of positive Darwinian evolution (Hughes and Nei, 1988, 1989). Since the diversification of drug-metabolizing P450 genes in animals seems to have been promoted through adaptation to toxic materials in foods (Krieger *et al.*, 1971; Gonzalez and Nebert, 1990), the consequences of adaptive evolution might be traced in the nucleotide sequences of P450 genes as in major histocompatibility complex or immunoglobulin genes.

Keeping this possibility in mind, I analyzed the coding nucleotide sequences of P450 family 2 (CYP2) genes in detail and found that the regions encoded by codons with excessively high nonsynonymous substitutions coincided well with the potential substrate-binding sites inferred from the refined alignment between CYP2 and bacterial P450 protein sequences. The CYP2 family was chosen for study because it consists of many related genes that are suitable for analysis of nucleotide substitution patterns and because there have been many experimental studies on members of this family. I show here that all the known point mutations and chimeric fragments that alter substrate specificities of some members of the CYP2 family are located within or very close to the putative substrate recognition sites identified by comparative sequence analysis methods.

MATERIALS AND METHODS

Nucleotide and Amino Acid Sequence Data—All coding nucleotide sequences were taken directly from the literature and cross-checked with the data in GenBank Rel. 66 when possible. Two sequences were regarded as polymorphic when their derived amino acid sequences differed by less than 1%, and the sequence that was complete or closest to the consensus sequence was selected. Thus, 51 CYP2 (2A1-7, 2B1-5, 7, 9, 10, 13 with two entries for each 2B4 and 2B5, 2C1-9, 11-14, 16, 18, 19, 21, 22, 2D1-6, 9, 10, three mammalian 2E1, 2F1, 2G1, 2H1, and 2H2), 12 CYP1 (five mammalian 1A1, six mammalian 1A2, and one fish 1A1), eight CYP3A (3A1-8), and eight bacterial (101A, 103A, 104A, 105A-C, 106A, and 55A) sequences were used for protein sequence analysis. *Fusarium oxysporum* P450 55A was included in the bacterial group because this gene was probably integrated into the fungal genome via a recent horizontal gene transfer event from a bacterium related to *Streptomyces* (Kizawa *et al.*, 1991). The codon substitution patterns of the sequences of CYP2A, 2B, 2C, and 2D subfamily members were analyzed.

Alignment of Groups of Protein Sequences—CYP1, 2, and 3 and bacterial amino acid sequences were aligned within each group by the methods described previously (Gotoh and Fujii-Kuriyama, 1989; Gotoh, 1990). The sequences of two groups were aligned using a few modifications of the standard dynamic programming algorithm with linear gap weights (Gotoh, 1982). First, a column in the original alignment was converted asymmetrically to either the "frequency" or "profile" vector, as described by Gribskov *et al.* (1987). In calculating a frequency or profile, we used a set of weights to even the contributions of individual sequences (Gotoh and Fujii-Kuriyama, 1989; Altschul *et al.*, 1989). Second, we introduced four additional elements in a frequency or profile vector, *i.e.* helix-, sheet-, and coil-forming propensities calculated by the method of Gibrat *et al.* (1987) and a hydrophathy index obtained with the parameters of Kyte and Doolittle (1982) with a window size of 9. A similarity score used in the alignment process was the scalar product of frequency and profile vectors. The contributions of the three parts, *i.e.* primary structure, secondary structure, and hydrophathy, may be changed with normalizing factors. In the present analysis, we adjusted these factors to 0.5:0.25:0.25, where each factor stands for the portion of a score

coming from the corresponding part for two identical random sequences with the average amino acid compositions. The value for each element is an average of those computed for individual sequences, in which we used the same weights as those in calculating the profiles.

Analysis of Nucleotide Substitution Patterns—The coding nucleotide sequences of CYP2A, 2B, 2C, and 2D members were aligned within each subfamily according to protein sequence alignments. The few gaps found in the alignments were ignored in the following analysis. The numbers of synonymous, $N_s(A, B)$, and nonsynonymous, $N_n(A, B)$, sites and the numbers of synonymous, $n_s(A, B)$, and nonsynonymous, $n_n(A, B)$, substitutions between a pair of sequences, A and B, were estimated by the method of Nei and Gojobori (1986). The fractions of synonymous and nonsynonymous substitutions are defined as $p_s = n_s(A, B)/N_s(A, B)$ and $p_n = n_n(A, B)/N_n(A, B)$, respectively. Let $n_{si}(A, B)$ and $n_{ni}(A, B)$ be the numbers of synonymous and nonsynonymous substitutions between sequence A and B within a window of the size of w codons centered at the i th codon position ($w/2 \leq i \leq L - w/2$, where L is the total number of codons in A or B). $R_{si}(A, B) = n_{si}(A, B) \cdot L/n_s(A, B) \cdot w$ and $R_{ni}(A, B) = n_{ni}(A, B) \cdot L/n_n(A, B) \cdot w$ indicate local variations in synonymous and nonsynonymous substitutions normalized with the expected numbers based on the global values. Assuming that R_{si} and R_{ni} are intrinsic to the codon position i , regardless of the sequence pairs considered, we estimated the most plausible value for R_{si} or R_{ni} by linear regression analysis, *i.e.* R_{si} (or R_{ni}) was estimated from the regression coefficient of n_{si} (or n_{ni}) on $n_s \cdot w/L$ (or $n_n \cdot w/L$) for various pairs of sequences. Since n_{si} and n_{ni} tend to become saturated as the sequence divergence increases, we considered only pairs for which $(n_s + n_n)/L \leq 0.17$, *i.e.* overall nucleotide sequence divergences are less than 17%. This eliminated all interspecies comparisons, except a few between rodent sequences. We also imposed weights to individual pairs, obtained according to Model I of Altschul *et al.* (1989), to correct for otherwise excessive contributions of phylogenetically differentiated pairs.

RESULTS

Structural Conservation between CYP2 and Bacterial P450s—Fifty-one CYP2 sequences and eight bacterial sequences were aligned with each other by the group-to-group alignment method as described under "Materials and Methods." The result is illustrated in Fig. 1, in which the profiles of secondary structure information (a-c) and hydrophathy indices (d) averaged over CYP2 proteins (*upper profile* in each panel) are juxtaposed together with those for bacterial P450s (*lower profile*). A horizontal flat line in a profile indicates the location of a gap (a span of deleted residues) introduced in the alignment process. The longest gaps are located between *Helices F* and *G* and between *Helices J* and *K*. Except for these gaps and about 30 positions at the N terminus, the CYP2 profiles clearly resemble the bacterial profiles. The dissimilarity in the N-terminal region is reasonable because this region is decisive for the difference in subcellular localizations of membrane-bound microsomal P450s and soluble bacterial P450s (Sakaguchi *et al.*, 1987; Vergères *et al.*, 1989).

To evaluate the alignment more quantitatively, I examined how the secondary structure or hydrophathy index at each site in CYP2 correlates with that for bacterial P450s at the same alignment position. As references for this comparison, several pairs of P450 groups with various degrees of sequence divergence were examined similarly, and results are summarized in Fig. 2. All but one of the correlation coefficients exceeded 0.5. Even the weakest correlation between the β -structure indices of CYP2 and bacterial P450s ($r(\text{sheet}) = 0.42$) is highly significant ($t = 9.4$, $p < 10^{-16}$). The correlation coefficients gradually decline as sequence divergence increases. For all comparisons, a definite order was observed among the four correlation coefficients: *i.e.* $r(\text{hydrophathy}) > r(\text{coil}) > r(\text{helix}) > r(\text{sheet})$. The absence of abnormality in the CYP2-bacterial comparisons suggests that the quality of the alignment is reasonably good and is similar to those between mammalian P450 families.

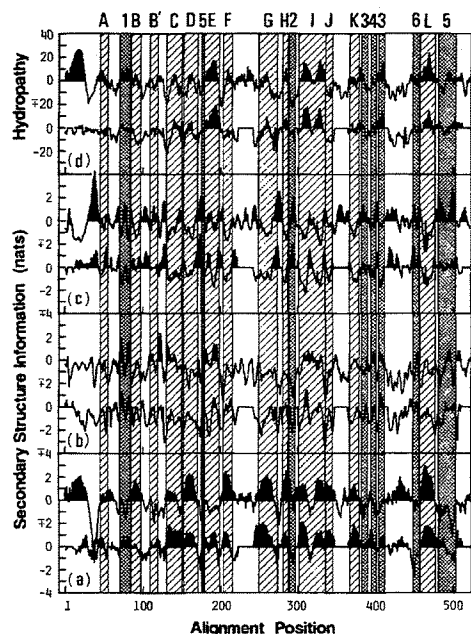


FIG. 1. Predicted secondary structure information and hydropathy index calculated for collective CYP2 sequences and bacterial sequences. Helix- (a), sheet- (b), and coil- (c) forming propensities determined by the method of Gibrat *et al.* (1987). d, hydropathy profiles obtained with the method of Kyte and Doolittle (1982). In each panel, the upper profile for CYP2 sequences and the lower profile for bacterial sequences are juxtaposed according to the optimal alignment between the two groups of sequences. The hatched and cross-hatched areas indicate the locations of α -helices and β -structures, respectively, in P450 101A (Poulos *et al.*, 1987). The labels above the panels are according to Poulos *et al.* (1987), except for the Cys ligand loop, which is labeled 6.

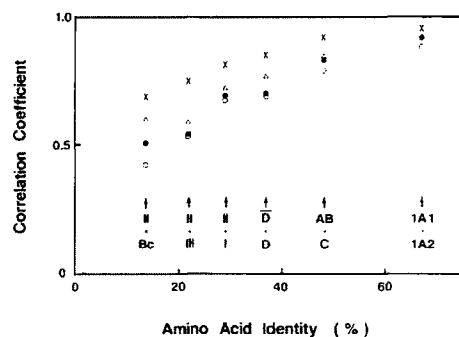


FIG. 2. Correlation coefficients of secondary structure information and hydropathy indices between various groups of P450 sequences. Correlation coefficients were calculated from the values for helix (●), sheet (○), or coil (△) information and the hydropathy index (×) at the same positions in alignments between various pairs of P450 groups. I, II, and III, the CYP1, 2, and 3 families, respectively; AB, combined sequences of subfamilies 2A and 2B; C, the 2C subfamily; D, the 2D subfamily; D, the CYP2 sequences other than those of 2D; Bc, bacterial sequences closely related to 101A. The abscissa indicates the average percentage of amino acid identities between the two sequence groups compared.

The hatched areas in Fig. 1 indicate the positions corresponding to the 13 helices in P450 101A (Poulos *et al.*, 1987), and the cross-hatched areas indicate those corresponding to β -structures. The labels at the top of the figure of the α -helical and β -sheet regions are according to the notation of Poulos *et al.* (1987), except the Cys ligand loop, which is labeled 6. Most of the tentative helical regions in both CYP2 and bacterial sequences have strong helix-forming propensities. Of 210 sites

at which P450 101A residues have a helical conformation, 141 (67%), 150 (71%), and 146 (70%) sites are predicted to be α -helical with the original P450 101A sequence, collective bacterial sequences, and collective CYP2 sequences, respectively, when the decision constants in each case were adjusted so that the predicted total contents of helices and sheets fit those of the three-dimensional structure of P450 101A. Edwards *et al.* (1989) reported a similar degree of predictivity of α -helices using an older version of the method of Garnier *et al.* (1978). The accuracies of prediction for β -structures were worse, being 44%, 40%, and 35% for the three sets of sequences. This poor prediction is probably related to the facts that β -structures are rare in P450 proteins, and β -structure information was least conserved in different groups of P450s (Fig. 2). The cumulative accuracies for the three states (helix, β -structure, and coil) were 59%, 60%, and 56%, respectively, for the three sequence sets. These values are close to the mean (58%) for the whole data base reported by Gibrat *et al.* (1987). The most important finding here was that nearly the same degrees of prediction were attained with CYP2 sequences and with the original P450 101A sequence, although on the average these sequences share only 13% of identical amino acids. This fact strongly supports the idea that the three-dimensional structure of P450 101A is basically retained in CYP2 proteins, as suggested in previous reports (Fujii-Kuriyama *et al.*, 1987; Gotoh and Fujii-Kuriyama, 1989; Nelson and Strobel, 1989).

Alignment-based Prediction of Substrate Recognition Sites in CYP2 Proteins—X-ray crystallography of substrate-bound forms of P450 101A (Poulos *et al.*, 1985, 1987) showed that the substrate camphor interacts with protein residues dispersed in several separate loci in the primary structure. Based on the three-dimensional model, Laughton *et al.* (1990) listed the residues in P450 101A that are within 10 Å of the bound camphor molecule. Shaded areas in Fig. 3 indicate the locations of these substrate-binding residues on our alignment between CYP2 and bacterial P450s. The CYP2 members in Fig. 3 are selected to represent all the gaps within the CYP2 alignment, except those at the extreme N or C termini.

The encircled residues in Fig. 3 are taken from mutant or chimeric cDNA clones of P450 2A4/2A5 (Lindberg and Negishi, 1989), P450 2B2 (Aoyama *et al.*, 1989), P450 2C4/2C5 (Kronbach *et al.*, 1989), and P450 2D1 (Matsunaga *et al.*, 1990b) that show markedly different substrate specificities (rather than enzymatic activities) from the parental clones upon expression. These residues form three clusters, all of which lie within, or in the close vicinity of (not more than three amino acid residues apart from) the alignment positions corresponding to the substrate-binding sites of P450 101A (Fig. 3). These encircled residues are not the minimal set responsible for the altered substrate specificities. For example, amino acids changed at only one or two of the three encircled sites in the B'-C region might be enough for the different catalytic activities of P450 2C4 and P450 2C5 (Kronbach *et al.*, 1989).

Imai and co-workers (Imai, 1988; Uno and Imai, 1989; Uno *et al.*, 1990) constructed several chimeras between P450 2C2 and 2C14 and found three separate regions that are likely to interact with substrate molecules. The sites corresponding to these three regions are indicated by solid lines beneath the P450 2C4 sequence in Fig. 3. All three regions cover some of the alignment positions corresponding to the P450 101A substrate-binding residues. Although not shown in Fig. 3, residues in the distal helix (Helix I) are known to be involved in binding of substrates (Poulos *et al.*, 1985, 1987; Imai and Nakamura, 1988, 1989; Furuya *et al.*, 1989a, 1989b; Zhou *et al.*, 1991). Thus, the substrate recognition sites in CYP2

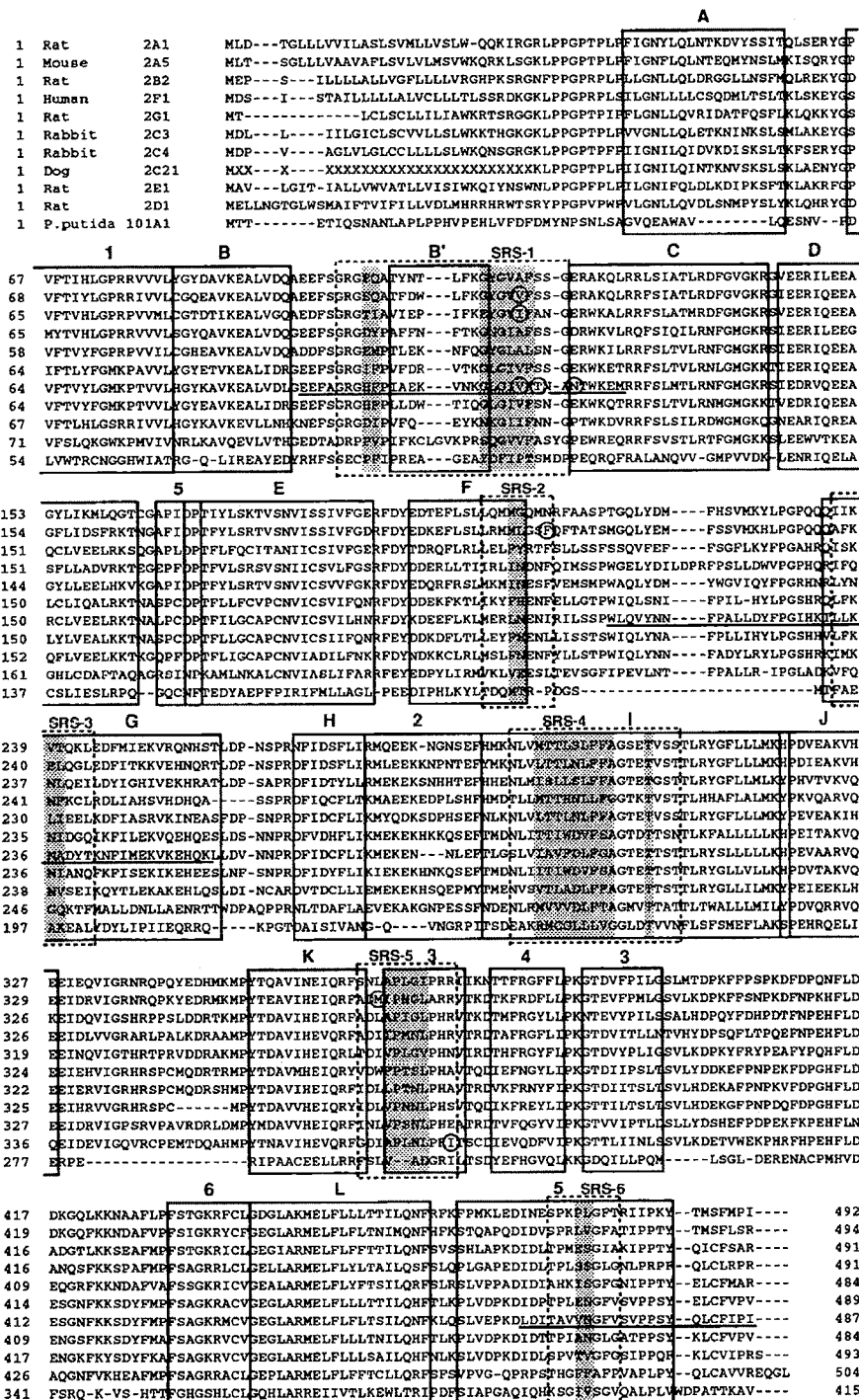


FIG. 3. Sequence alignment between representative CYP2 members and P450 101A. The CYP2 members were selected from a complete CYP2 alignment so that all gaps, except at the extreme N and C termini, should be presented. The regions corresponding to the helices and β -structures in P450 101A are boxed with solid lines with the same labels as in Fig. 1. Shaded areas indicate the positions corresponding to the substrate-binding residues in P450 101A (Laughton *et al.*, 1990). Six SRSs (see text) are boxed with broken lines. The residues identified experimentally as responsible for substrate specificity are circled. Underlines indicate the fragments that affect substrate specificities in chimeras between rabbit P450 2C2 and 2C14.

proteins suggested from the alignment are fully compatible with available experimental observations.

From all these observations, the following six separate regions were tentatively assigned as substrate recognition sites (SRSs) in CYP2 proteins: 1) B' and flanking areas (103–126), 2) the C-terminal end of Helix F (209–216), 3) the N-terminal end of Helix G (248–255), 4) the N-terminal half of Helix I (302–320), 5) the β 3 area (375–385), and 6) a central region of β 5 (485–493), where the numbers in parentheses refer to the position in the alignment shown in Fig. 3. These regions correspond to the P450 101A substrate-binding sites extended by three amino acid residues on both sides. Since

there are only a few residues between the first and second of these regions, these two regions were fused and the combined area was named SRS-1. SRSs account for 76–79 of the total 512 positions in the CYP2 alignment.

Nucleotide Substitution Patterns in Members of CYP2 Subfamilies—To confirm the assignment of substrate recognition sites in CYP2 proteins, I examined local sequence variability between duplicate members within a subfamily. The underlying hypothesis for this was that closely related drug-metabolizing P450s may show divergent substrate specificities and so cooperate in metabolizing a wider range of foreign compounds. If this is so, the amino acid sequences of

substrate recognition sites should be more variable than the rest of the molecule. To locate variable regions, I examined coding nucleotide sequences rather than amino acid sequences, since nucleotide sequences provide much information about the molecular evolutionary mechanisms operating in diversification of closely related genes.

Fig. 4 plots $\Delta R_i = R_{ni} - R_n$ at each codon position calculated for each of four CYP2 subfamilies. R_{ni} is the mean ratio of the real to expected numbers of nonsynonymous substitutions within a window centered at the codon position i , and R_n is a similar value for synonymous substitutions (see "Materials and Methods"). Thus if ΔR_i is positive, the region covered by the window has accumulated a larger number of nonsynonymous nucleotide substitutions than of synonymous substitutions, whereas if ΔR_i is negative, amino acid changes within the window are fewer than expected from the average nucleotide substitution rates. A good correlation between the regions with positive ΔR_i values and SRSs (Fig. 4, shaded areas) is apparent for all subfamilies. χ^2 tests with 2×2 contingency tables (Table I) confirmed the significant associations in all cases: $\chi^2 = 39.0, 37.7, 49.2$, and 28.9 ($\gg 10.8$, $p < 0.001$ with

1° of freedom) for 2A, 2B, 2C, and 2D, respectively. The correlation is particularly good for the 2C subfamily, presumably because this 2C subfamily contains the largest number of members and its evolutionary process accords well with our hypothetical scheme. As an exception, the ΔR_i values in SRS-5 (β 3 area) are not positive. Since this region is narrow and is flanked by well conserved residues such as Glu and Arg in Helix K and His or Arg in β 3 (Gotoh and Fujii-Kuriyama, 1989), amino acid changes in SRS-5 might be restricted compared with those in other SRSs.

Table II lists the fractions of synonymous nucleotide substitutions per synonymous site (p_s) and fractions of nonsynonymous substitutions per nonsynonymous site (p_n) between intraspecies P450 genes calculated separately within and outside SRSs. As expected, p_n/p_s values for SRSs are all larger than those for other regions. The values outside SRSs ($p_n/p_s = 0.30 \pm 0.08$) are fairly constant in all pairs and are similar to the values observed in various genes (Miyata *et al.*, 1980; Nei, 1987). (The unusually large value for rat 2D1 and 2D5 is probably due to the large scale gene conversion between these genes (Matsunaga *et al.*, 1990a) and hence is omitted from the present statistics.) On the other hand, the p_n/p_s values for SRSs vary extensively ($p_n/p_s = 0.74 \pm 0.35$). Although $p_n > p_s$ for some pairs, the inequality is not significant in any case.

DISCUSSION

We (Gotoh *et al.*, 1983; Gotoh and Fujii-Kuriyama, 1989) and others (Black and Coon, 1987; Kalb and Loper, 1988; Nelson and Strobel, 1987, 1988; Edwards *et al.*, 1989; Laughton *et al.*, 1990) have reported several versions of alignment between bacterial and eukaryotic (including CYP2) P450 sequences. These alignments are generally consistent in the C-terminal half of the sequences (Helix I to the C terminus) in which conserved elements common to all P450s are located (for review, see Gotoh and Fujii-Kuriyama, 1989). In the N-terminal half, the regions of Helices C, D, and G are relatively well conserved, and even the earliest version (Gotoh *et al.*, 1983) between single sequences (2B1 versus 101A) is basically the same as that shown in Fig. 3 in terms of alignment of these regions. In contrast, the presumable substrate-binding regions exhibit extensive sequence variations, and so it has been hard to obtain unequivocal alignment of these regions. However, we now have the following reasons for the global correctness of our alignment (Fig. 3), including the potential substrate recognition regions.

First, the profiles of secondary structure information and hydropathy indices obtained with bacterial and CYP2 sequences match each other very well (Fig. 1). I was not the first to include predicted secondary structure information into alignment of P450 sequences. Edwards *et al.* (1989) mainly used helix-forming propensities to align various families of P450 sequences. Their alignment differs from ours in several respects and particularly in the locations of Helices E and F. This is not surprising because the Helix E region shows

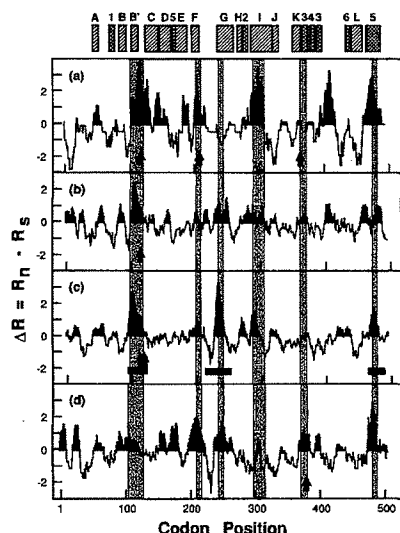


FIG. 4. Local variations in difference between rates of nonsynonymous nucleotide substitutions and synonymous substitutions. Positive ΔR_i means that nucleotide substitutions replacing amino acids within a window (nine codons) have occurred more frequently than expected from the global number of nonsynonymous substitutions after correction for local variation in nucleotide mutation rates. Calculations were made individually for four CYP2 subfamilies, 2A (a), 2B (b), 2C (c), and 2D (d). Shaded areas indicate SRSs. The locations of the residues identified experimentally as responsible for substrate recognition are indicated by arrowheads. The three filled boxes in c indicate the locations of the fragments that affect substrate specificities in chimeras between rabbit P450 2C2 and 2C14. The potential helical and β -structure regions are indicated above the panels by boxes. The same labels as in Figs. 1 and 3 are used.

TABLE I

Contingency tables for test of excess nonsynonymous substitutions in SRSs

The numbers of residues with positive (+) and negative (−) ΔR_i values (Fig. 4) within and outside SRSs are counted separately. χ^2 values were obtained with Yates correction.

ΔR	CYP2A ($\chi^2 = 39.0$)			CYP2B ($\chi^2 = 37.7$)			CYP2C ($\chi^2 = 49.2$)			CYP2D ($\chi^2 = 28.9$)		
	+	−	Total	+	−	Total	+	−	Total	+	−	Total
SRSs	57	19	76	61	15	76	61	15	76	59	20	79
Others	150	269	419	173	246	419	152	268	420	175	251	426
Total	207	288	495	234	261	495	213	283	496	234	271	505

TABLE II
Fractions of synonymous (p_s) and nonsynonymous (p_n) nucleotide substitutions outside and inside SRSs

Species	Genes	Whole p_r^a	Outside SRS			Inside SRS		
			p_n	p_s	p_n/p_s	p_n	p_s	p_n/p_s
Rat	2A1 2A2	7.2 ± 0.7	4.0 ± 0.6	12.7 ± 2.0	0.31	13.9 ± 2.6	12.2 ± 4.3	1.15
Mouse	2A4 2A5	1.5 ± 0.3	0.6 ± 0.3	2.8 ± 1.0	0.23	2.8 ± 1.2	5.1 ± 2.8	0.56
Rat	2B1 2B2	2.4 ± 0.4	1.1 ± 0.3	6.1 ± 1.4	0.17	3.4 ± 1.4	3.4 ± 2.4	0.98
Rat	2B1 2B3	16.6 ± 1.1	10.9 ± 1.0	29.0 ± 2.7	0.37	18.5 ± 2.9	41.1 ± 6.4	0.45
Rat	2B2 2B3	16.6 ± 1.1	11.1 ± 1.0	29.2 ± 2.7	0.38	18.3 ± 2.9	38.2 ± 6.4	0.48
Rabbit	2B4a 2B4b	1.3 ± 0.3	0.5 ± 0.2	3.9 ± 1.1	0.14	0.6 ± 0.6	1.7 ± 1.6	0.34
Rabbit	2B4a 2B5a	2.2 ± 0.4	0.8 ± 0.3	5.8 ± 1.3	0.13	2.8 ± 1.2	5.0 ± 2.8	0.56
Rabbit	2B4b 2B5a	2.2 ± 0.4	1.1 ± 0.3	4.8 ± 1.2	0.22	3.4 ± 1.4	3.3 ± 2.3	1.02
Rabbit	2B4a 2B5b	2.9 ± 0.5	1.3 ± 0.4	6.8 ± 1.4	0.19	2.8 ± 1.2	8.3 ± 3.6	0.34
Rabbit	2B4b 2B5b	2.6 ± 0.5	1.3 ± 0.4	5.5 ± 1.3	0.23	3.4 ± 1.4	6.7 ± 3.2	0.51
Rabbit	2B5a 2B5b	3.0 ± 0.5	1.5 ± 0.4	5.8 ± 1.3	0.26	3.4 ± 1.4	10.0 ± 3.9	0.34
Mouse	2B9 2B10	11.8 ± 0.9	6.4 ± 0.8	23.0 ± 2.5	0.28	16.8 ± 2.8	27.4 ± 5.8	0.61
Human	2C8 2C9	15.1 ± 1.0	9.6 ± 0.9	27.6 ± 2.7	0.35	20.9 ± 3.0	28.7 ± 6.0	0.73
Human	2C8 2C18	15.4 ± 1.0	9.6 ± 0.9	27.2 ± 2.7	0.35	23.9 ± 3.2	23.1 ± 6.0	0.85
Human	2C9 2C18	13.0 ± 1.0	7.7 ± 0.9	23.9 ± 2.6	0.32	19.2 ± 2.9	29.3 ± 6.2	0.66
Human	2C8 2C19	15.5 ± 1.0	9.8 ± 1.0	28.9 ± 2.7	0.34	21.5 ± 3.1	26.9 ± 5.8	0.80
Human	2C9 2C19	5.2 ± 0.6	3.6 ± 0.6	9.0 ± 1.7	0.40	5.2 ± 1.6	13.6 ± 4.6	0.39
Human	2C18 2C19	13.4 ± 1.0	7.9 ± 0.9	24.5 ± 2.6	0.32	20.9 ± 3.0	29.2 ± 6.1	0.71
Rat	2C12 2C13	11.7 ± 0.9	7.3 ± 0.8	18.1 ± 2.3	0.40	21.9 ± 3.1	22.1 ± 5.6	0.99
Rabbit	2C1 2C2	11.6 ± 0.9	6.3 ± 0.8	23.7 ± 2.5	0.27	17.6 ± 2.8	19.9 ± 5.3	0.88
Rabbit	2C1 2C14	10.3 ± 0.9	6.6 ± 0.8	17.7 ± 2.3	0.37	18.0 ± 2.9	9.8 ± 4.0	1.83
Rabbit	2C2 2C14	13.7 ± 1.0	8.2 ± 0.9	27.3 ± 2.6	0.30	19.1 ± 2.9	20.5 ± 5.5	0.93
Rabbit	2C4 2C5	3.3 ± 0.5	1.5 ± 0.4	5.3 ± 1.3	0.28	6.9 ± 1.9	11.4 ± 4.2	0.61
Rabbit	2C4 2C16	6.5 ± 0.7	3.0 ± 0.6	12.8 ± 2.0	0.23	11.7 ± 2.4	17.4 ± 5.0	0.67
Rabbit	2C5 2C16	5.9 ± 0.7	2.5 ± 0.5	11.8 ± 1.9	0.21	11.1 ± 2.3	17.7 ± 5.1	0.63
Rat	2D1 2D3	15.7 ± 1.0	9.3 ± 0.9	24.0 ± 2.4	0.39	24.8 ± 3.2	47.8 ± 6.6	0.52
Rat	2D1 2D5	1.9 ± 0.4	1.6 ± 0.4	1.5 ± 0.7	1.10	4.0 ± 1.5	1.6 ± 1.6	2.41
Rat	2D3 2D5	15.7 ± 1.0	9.3 ± 0.9	23.1 ± 2.4	0.40	27.2 ± 3.3	45.2 ± 6.5	0.60
Mouse	2D9 2D10	7.8 ± 0.8	5.4 ± 0.7	12.9 ± 1.9	0.42	12.1 ± 2.5	7.9 ± 3.4	1.52

^a Overall nucleotide sequence divergence within coding regions.

negative helix propensities in the CYP2 profile (Fig. 1a). However, the low helix propensities in this region are reasonable, because some CYP2 members, as well as P450 101A contain one or two prolines in the Helix E region. Our alignment is supported by the nearly identical shapes in other two secondary structure profiles (Fig. 1, b and c) and especially in the hydropathy profile (Fig. 1d). As shown in Fig. 2, hydropathy and coil propensity are conserved better than the helix propensity in all P450 families examined. Hence the use of helix propensity alone is less reliable than our method that considers various types of information.

The second point that supports our alignment is the scarcity of gaps within the suggested helical or β -structure regions in the alignments of all CYP2 members. It is noteworthy that the sequences in Fig. 3 represent all the gaps within the CYP2 alignment, except those located at the two ends. CYP2D members have 3 additional residues and CYP2E members have 1 less residue in the Helix B' region than other CYP2 members. Helix B' was not detected in P450 101A by x-ray crystallography at 2.6-Å resolution (Poulos *et al.*, 1985) but was seen in the 1.63-Å refined structure (Poulos *et al.*, 1987). It is not surprising that the regular secondary structure in this region is not conserved in other P450 proteins, since this B' region is one of the major substrate recognition sites (Figs. 3 and 4). The deletions found in the β 2 region of some CYP2 members are likely to occur in the turn or loop that connects opposite strands in the body of the antiparallel β -structure. Besides the gaps in the B' and β 2 regions, only two single-residue deletions are found near an end of Helix G and of β 5. The gaps in Helices A and B in the 101A sequence were introduced in the course of alignment of bacterial sequences, indicating somewhat variable natures of these short helices. Otherwise, the distributions of gaps in both 101A and CYP2 sequences are well consistent with the general tendency for

deletion or insertion of residues to occur outside secondary structures (Lesk *et al.*, 1986).

The third and most important point is the fact that our alignment accords nearly perfectly with experimentally identified substrate recognition sites in various CYP2 members. As shown in Fig. 3, all the known point mutations and chimeric fragments that significantly affect substrate specificities are mapped closely to the alignment-based SRSs. The importance of SRS-1 (B'-C area) for binding substrates has been repeatedly noted (Kronbach *et al.*, 1989; Uno and Imai, 1989; Aoyama *et al.*, 1989) and well explained by extrapolation from the three-dimensional structure of the camphor-bound form of P450 101A (Poulos *et al.*, 1985, 1987). Similarly, an important role of a part of the distal helix (SRS-4) in binding of substrates has been well documented (Poulos *et al.*, 1985, 1987; Imai and Nakamura, 1988, 1989; Furuya *et al.*, 1989a, 1989b; Zhou *et al.*, 1991). There are, however, few reports of structure-based interpretations of the other sites. For example, the structural basis of the critical amino acid residue 209 that determines testosterone hydroxylase activity of mouse P450 2A4 or coumarin 7-hydroxylase activity of 2A5 (Lindberg and Negishi, 1989) has been enigmatic (Iwasaki *et al.*, 1991). Now this residue is mapped in SRS-2 (F-G interhelical region) close to the C terminus of Helix F. The region spanning residues 211-262, which is essential for rabbit P450 2C2 (laurate ω -1 hydroxylase) to bind fatty acids (Imai, 1988), covers another SRS (SRS-3) located at the N terminus of Helix G. The flexibility of the F-G region of P450 101A monitored by temperature factors markedly reduces upon substrate binding (Poulos *et al.*, 1986). Hence the region was thought to be primarily important for accommodation of substrate molecules (Gotoh and Fujii-Kuriyama, 1989), although the corresponding region in eukaryotic sequences was only roughly assigned in our earlier report. The F-G inter-

helical region is the longest insertion in CYP2 sequences compared with bacterial ones. It is likely that this enlarged region facilitates accommodation of such large molecules as polycyclic hydrocarbons and steroids. A single amino acid substitution (Ile³⁸⁰ → Phe) in rat P450 2D1 resulted in decreased catalytic activity toward bufuralol but not debrisoquine (Matsunaga *et al.*, 1990b). This residue is mapped in SRS-5 (β 3 area), where one of the 3 residues responsible for the altered substrate specificities of mouse 2A4 and 2A5 (Lindberg and Negishi, 1989) is also located. Uno *et al.* (1990) noted that replacement of the C-terminal 28 residues of P450 2C2 with those of 2C14 produced a new stereospecific activity toward testosterone, suggesting participation of this region in substrate recognition. This C-terminal area overlaps SRS-6 within the β 5-structure. All these observations not only confirm our assignment of SRSs, but also afford structural and functional bases for our alignment.

Our assignment of SRSs is further supported by the results of analyses of local variations in nucleotide substitution patterns. Most interestingly, most of the main peaks in ΔR_i plots fall within SRSs (Fig. 4), and the associations between SRSs and positive ΔR_i values are highly significant for each of four CYP2 subfamilies (Table I). This strong association implies the functional importance of high amino acid replacement rates within SRSs and supports our initial hypothesis that duplicate genes coding for drug-metabolizing P450 enzymes with divergent substrate specificities are evolutionarily advantageous. In accordance with this hypothesis, there have been several reports indicating relationships between overall P450 activities and food habits (Krieger *et al.*, 1971; Ronis and Hodgson, 1989).

In the case of highly polymorphic human or mouse major histocompatibility complex genes, nonsynonymous substitution rates within the antigen recognition site were higher than synonymous rates, providing strong evidence for adaptive overdominant selection (Hughes and Nei, 1988, 1989). As listed in Table II, nonsynonymous rates within SRSs of CYP2 genes are slightly lower than synonymous rates. The average ratio of $p_n/p_s = 0.74 \pm 0.35$ does not by itself indicate whether the higher nonsynonymous rates within SRSs than those outside SRSs is due to neutral mutations or adaptive diversifications. The latter possibility is the more likely from circumstantial evidence, but further analyses, preferably with a larger set of sequence data, may be required to draw a conclusion on the molecular evolutionary mechanism.

I also examined nucleotide substitution patterns of three other drug-metabolizing P450 families (data not shown). Nucleotide changes between two members of the CYP1 family, 1A1 and 1A2, are nearly saturated, giving rise to a featureless ΔR_i pattern. The substitution pattern of the CYP3A subfamily was basically similar to those of CYP2 subfamilies (Fig. 4), suggesting that CYP2 and CYP3 have undergone common evolutionary processes. On the other hand, the pattern for CYP4A was quite different from those of CYP2 or CYP3. The functional and/or structural reasons for this remain to be elucidated.

In summary, several independent lines of evidence indicate that substrate recognition regions in CYP2 proteins are dispersed along the primary structure. This paper reports six such regions, SRS-1-6, which constitute about 16% of the total residues in the P450 molecule. It is very likely that corresponding regions in other eukaryotic families of P450 are involved in binding specific substrates, although the six SRSs may differ in relative importance in different families or subfamilies. Most of the residues that participate in substrate binding are probably included in these six SRSs, al-

though a few sites may be missing from the present list. The present findings will be useful for molecular design of engineered P450 enzymes with new substrate specificities.

Acknowledgments—I am grateful to Dr. Masatoshi Nei for support and advice on the Nei and Gojobori method during my stay in his laboratory. I also thank Etsuko Yomo for secretarial assistance.

REFERENCES

- Altschul, S. F., Carroll, R. J., and Lipman, D. J. (1989) *J. Mol. Biol.* **207**, 647-653
- Aoyama, T., Korzekwa, K., Nagata, K., Adesnik, M., Reiss, A., Lapenson, D. P., Gillette, J., Gelboin, H. V., Waxman, D. J., and Gonzalez, F. J. (1989) *J. Biol. Chem.* **264**, 21327-21333
- Barton, G. J., and Sternberg, M. J. E. (1987) *J. Mol. Biol.* **198**, 327-337
- Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. (1987) *Nature* **329**, 512-518
- Black, S. D., and Coon, M. J. (1987) *Adv. Enzymol.* **60**, 35-87
- Edwards, R. J., Murray, B. P., Boobis, A. R., and Davies, D. S. (1989) *Biochemistry* **28**, 3762-3770
- Fujii-Kuriyama, Y., Sogawa, K., and Gotoh, O. (1987) in *Pharmacology* (Rand, M. J., and Raper, C., eds) pp. 771-774, Elsevier, Amsterdam
- Furuya, H., Shimizu, T., Hatano, M., and Fujii-Kuriyama, Y. (1989a) *Biochem. Biophys. Res. Commun.* **160**, 669-676
- Furuya, H., Shimizu, T., Hirano, K., Hatano, M., Fujii-Kuriyama, Y., Raag, R., and Poulos, T. L. (1989b) *Biochemistry* **28**, 6848-6857
- Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) *J. Mol. Biol.* **120**, 97-120
- Gibrat, J.-F., Garnier, J., and Robson, B. (1987) *J. Mol. Biol.* **198**, 425-443
- Gonzalez, F. J. (1990) *Pharmacol. Ther.* **45**, 1-38
- Gonzalez, F. J., and Nebert, D. W. (1990) *Trends Genet.* **6**, 182-186
- Gotoh, O. (1982) *J. Mol. Biol.* **162**, 705-708
- Gotoh, O. (1990) *Bull. Math. Biol.* **52**, 359-373
- Gotoh, O., and Fujii-Kuriyama, Y. (1989) in *Frontiers in Biotransformation* (Ruckpaul, K., and Rein, H., eds) Vol. 1, pp. 195-243, Akademie-Verlag, Berlin
- Gotoh, O., Tagashira, Y., Iizuka, T., and Fujii-Kuriyama, Y. (1983) *J. Biochem. (Tokyo)* **93**, 807-817
- Gotoh, O., Tagashira, Y., Morohashi, K., and Fujii-Kuriyama, Y. (1985) *FEBS Lett.* **188**, 8-10
- Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84**, 4355-4358
- Hughes, A. L., and Nei, M. (1988) *Nature* **335**, 167-170
- Hughes, A. L., and Nei, M. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 958-962
- Imai, Y. (1988) *J. Biochem. (Tokyo)* **103**, 143-148
- Imai, Y., and Nakamura, M. (1988) *FEBS Lett.* **234**, 313-315
- Imai, Y., and Nakamura, M. (1989) *Biochem. Biophys. Res. Commun.* **158**, 717-722
- Iwasaki, M., Juvonen, R., Lindberg, R., and Negishi, M. (1991) *J. Biol. Chem.* **266**, 3380-3382
- Kalib, V. F., and Loper, J. C. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7221-7225
- Kizawa, H., Tomura, D., Oda, M., Fukamizu, A., Hoshino, T., Gotoh, O., Yasui, T., and Shoun, H. (1991) *J. Biol. Chem.* **266**, 10632-10637
- Krieger, R. I., Feeny, P. P., and Wilkinson, C. F. (1971) *Science* **172**, 579-581
- Kronbach, T., Larabee, T. M., and Johnson, E. F. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 8262-8265
- Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105-132
- Laughton, C. A., Neidle, S., Zvelebil, M. J. J. M., and Sternberg, M. J. E. (1990) *Biochem. Biophys. Res. Commun.* **171**, 1160-1167
- Lesk, A. M., Levitt, M., and Chothia, C. (1986) *Protein Eng.* **1**, 77-78
- Lindberg, R. L. P., and Negishi, M. (1989) *Nature* **339**, 632-634
- Matsunaga, E., Umeno, M., and Gonzalez, F. J. (1990a) *J. Mol. Evol.* **30**, 155-169
- Matsunaga, E., Zeugin, T., Zanger, U. M., Aoyama, T., Meyer, U. A., and Gonzalez, F. J. (1990b) *J. Biol. Chem.* **265**, 17197-17201
- Miyata, T., Yasunaga, T., and Nishida, T. (1980) *Proc. Natl. Acad. Sci. U. S. A.* **77**, 7328-7332
- Nebert, D. W., Nelson, D. R., Coon, M. J., Estabrook, R. W., Feyereisen, R., Fujii-Kuriyama, Y., Gonzalez, F. J., Guengerich, F. P.,

- Gunsalus, I. C., Johnson, E. F., Loper, J. C., Sato, R., Waterman, M. R., and Waxman, D. J. (1991) *DNA Cell Biol.* **10**, 1-14
- Nei, M. (1987) *Molecular Evolutionary Genetics*, pp. 64-110, Columbia University, New York
- Nei, M., and Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418-426
- Nelson, D. R., and Strobel, H. W. (1987) *Mol. Biol. Evol.* **4**, 572-593
- Nelson, D. R., and Strobel, H. W. (1988) *J. Biol. Chem.* **263**, 6038-6050
- Nelson, D. R., and Strobel, H. W. (1989) *Biochemistry* **28**, 656-660
- Onoda, M., Haniu, M., Yanagibashi, K., Sweet, F., Shively, J. E., and Hall, P. F. (1987) *Biochemistry* **26**, 657-662
- Picado-Leonard, J., and Miller, W. L. (1988) *Mol. Endocrinol.* **2**, 1145-1150
- Pompon, D., and Nicolas, A. (1989) *Gene (Amst.)* **83**, 15-24
- Poulos, T. L., Finzel, B. C., Gunsalus, I. C., Wagner G. C., and Kraut, J. (1985) *J. Biol. Chem.* **260**, 16122-16130
- Poulos, T. L., Finzel, B. C., and Howard, A. J. (1986) *Biochemistry* **25**, 5314-5322
- Poulos, T. L., Finzel, B. C., and Howard, A. J. (1987) *J. Mol. Biol.* **195**, 687-700
- Ronis, M. J. J., and Hodgson, E. (1989) *Xenobiotica* **19**, 1077-1092
- Sakaguchi, M., Mihara, K., and Sato, R. (1987) *EMBO J.* **6**, 2425-2431
- Sakaki, T., Shibata, M., Yabusaki, Y., and Ohkawa, H. (1987) *DNA (N. Y.)* **6**, 31-39
- Tanaka, T., and Nei, M. (1989) *Mol. Biol. Evol.* **6**, 447-459
- Uno, T., and Imai, Y. (1989) *J. Biochem. (Tokyo)* **106**, 569-574
- Uno, T., Yokota, H., and Imai, Y. (1990) *Biochem. Biophys. Res. Commun.* **167**, 498-503
- Vergères, G., Winterhalter, K. H., and Richter, C. (1989) *Biochemistry* **28**, 3650-3655
- Zhou, D., Pompon, D., and Chen, S. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88**, 410-414